Camera Modeling Technique of 3D Sensing Based on Tile Coding for Computer Vision

Toshihiko Watanabe
Osaka Electro-Communication University
18-8 Hatsu-cho, Neyagawa
Osaka, 572-8530, Japan
+81-72-824-1131
t-wata@isc.osakac.ac.jp

Yuichi Saito
Osaka Electro-Communication University
18-8 Hatsu-cho, Neyagawa
Osaka, 572-8530, Japan
+81-72-824-1131
mf11a005@oecu.jp

ABSTRACT
Recently, the 3D sensing technique using multiple cameras has been applied to various areas such as visualization, motion capturing, and so on. However, improvement of the camera model calibration is indispensible for more precise measurement. In this study, we propose a camera modeling technique for 3D sensing based on tile coding (CMAC) structure. A distance between a sensing target and the camera is used to construct the camera model considering optical projection characteristics. In our approach, the least mean square error method is successfully applied considering the simple tile structure to formulate the camera model. Then iterative calculations for solving the inverse problem of the 3-D to 2-D projection by camera are performed to attain measured 3D coordinates. Through sensing experiments of stereo vision measurement based on the proposed approach, we showed the performance of the model was drastically improved compared with the conventional modeling approach such as Open CV model or crisp partitioned model.

Keywords
Computer vision, Stereo Vision, Camera Model, Tile Coding, CMAC, Perspective Projection.

1. INTRODUCTION
Recently, sensing techniques using optical systems have been widely applied to various areas. The computer vision technique that can not only obtain a target measurement based on image processing and signal processing using cameras or projectors, but also construct a 3-Dimensional shape of the target in the computer, becomes a necessary technique for inspection of industrial products such as LSI (Large Scale Integration) and circuit board, measurement and archives of historical constructions, robotic vision, and inspection of large-sized industrial products such as motor vehicles.

BODYNETS 2013, September 30-October 02, Boston, United States

DOI 10.4108/icst.bodynets.2013.253708

In order to apply the computer vision techniques using cameras and/or projectors, it is indispensable to construct a precise "camera model" through calibration process. The camera model expresses the necessary relationship between 2D image coordinates and 3D world coordinates. Since the precision of the camera model directly affects the measurement precision, we need an accurate and robust model construction technique, in order to improve the performance of computer vision systems. However, the conventional model based on the pinhole model, i.e., perspective view model, is sometimes deteriorated caused by the error caused by approximation of the camera. Moreover, a lens distortion sometimes affects the precision of the camera model.

In this study, we propose a practicable modeling technique based on tile coding (CMAC) to construct a camera model. Our motivation of this study is to develop a more precise and robust camera model. The camera model is utilized based on stereo vision configuration of the sensing system. From empirical study, we utilize a distance between the target and cameras for modeling. Our former study [9] shows that the concept of the fuzzy model is promising for camera model construction to determine the parameters of the 3D to 2D projection. However, the practical performance in stereo vision was not confirmed because of the difficulty of inverse problems. In this paper, we propose simple and more practicable model structure based on tile coding and apply to the measurement problem of stereo vision. Through experiments of measurement in stereo vision using cameras for industry, we evaluate the proposed approach.

The remainder of the paper is constructed as follows. In section II, the 3D sensing technique is overviewed. The camera modeling based on tile coding for 3D sensing is proposed in the section III. The experimental results are presented in section IV. Finally, conclusions are drawn in section V.

2. 3D-SENSIG TECHNIQUE
In a measurement system based on cameras, the camera model [1-3] is indispensable to represent the transformation process that transforms 3D world coordinates of the target into 2D image coordinates. In other words, we should describe the physical phenomena mathematically how images in the camera reflect the real-world target.

2.1 Camera Model
In general, the mathematical model of camera for sensing is considered approximating as a pinhole camera. In the actual pinhole camera, an image reflects a target on the image plane. In the image plane, the image is always in focus and the size of image depends only on the focal distance and its distance from the camera. From these reasons, we generally assume the camera can

be modeled as the pinhole camera model for sensing problem. Figure 1 shows the pinhole camera model. In the pinhole camera model, based on the similar triangles, we have the following formulation.

$$x = f \cdot (X / Z) \tag{1}$$

where f is the focal distance, x is the size in the image, X is the actual size of the target, and Z denotes the actual distance between the camera and the target.

2.2 Calibration for Camera Model

The calibration for a camera model is an important step in constructing the camera model. Firstly, data pairs of two dimensional image data and corresponding three dimensional target coordinates through the camera are collected using calibrator such as the checkerboard. Then the necessary parameters comprising five internal parameters such as the focal distance and six external parameters such as direction and rotation information of the camera are estimated using collected data sets. All the parameters are called camera parameters. Figure 2 shows the relationship among the world coordinates, the camera coordinates, and the image coordinates. After decision of the camera parameters, the relationship between the coordinates is formulated. The detail of these parameters is described below.

2.3 Perspective Camera Model

The relationship between image coordinates and the target coordinates of pinhole camera is expressed using the perspective projection matrix based on internal parameters and external parameters such as geometric relations of camera position and direction. The perspective projection matrix P is defined by multiplication of internal parameters and external parameters as follows [1, 2]:

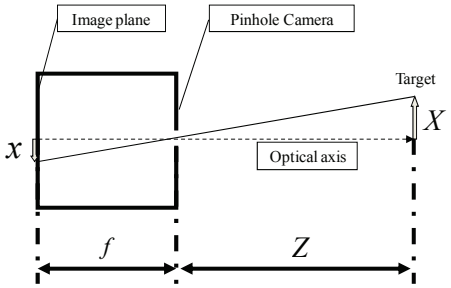

Figure 1. Pinhole camera model

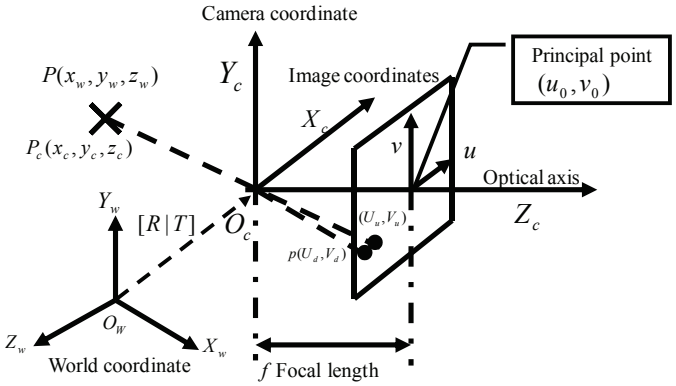

Figure 2. Transformation of coordinates

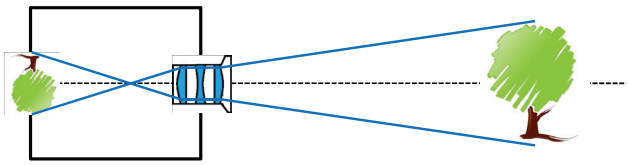

Figure 3. Conceptual figure of projection model of a camera with a contemporary lens

$$P = A[R \mid T] \tag{2}$$

where A is the internal parameter matrix, R is the rotation matrix, and T is the position vector of the camera. The internal parameter matrix A is defined as:

$$A = \begin{bmatrix} fk_u & fk_s & u_0 \\ 0 & fk_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

where f is the focal distance, (u_0, v_0) is the image center, k_u and k_v are the number of pixels per unit distance in image coordinates, i.e., abscissa and ordinate axes in the image, respectively, and k_s is a skew parameter. From these transformations, Eq.(1) is extended to the following essential formulation.

$$\omega \cdot U = P \cdot Q \tag{4}$$

$$U = [u \;\; v \;\; 1]^T \;,\; Q = [X \;\; Y \;\; Z \;\; 1]^T \tag{5}$$

where (u, v) is the image coordinates, (X, Y, Z) is the corresponding world coordinates, and ω is the scaling parameter. It should be noted that U and Q are represented as homogeneous coordinates. It should be also noted that the essential relationship is not linear according to the parameters because of the homogeneous representation.

In the conventional approach, the above parameters, ω and P, are estimated by some optimization techniques using pairs of target coordinates and corresponding image coordinates collected through the calibration process.

3. CAMERA MODELING TECHNIQUE BASED ON TILE CODING

In general, the camera model is assumed as Eq.(4) and Eq.(5). In the general approach of camera calibration, the internal matrix is estimated by Zhang method [4, 5] and the external matrix is estimated based on the fundamental technique such as PnP (Perspective n-Points) problem [2]. However, since the actual camera is not pinhole camera, there has been a problem of model precision substantially [6]. The lens distortion and variation of manufacturing such as optical axis variation sometimes affect model precision. Fig.3 shows the conceptual figure of the necessary projection model of a camera. Obviously, modern cameras used in the computer vision applications are not a pinhole camera and the conventional model (perspective projection) based on the pinhole camera includes approximation errors especially in the depth direction. We also confirmed the fact through sensing experiments that the model precision deteriorates in the change of distance in the depth direction. For these reasons, in order to improve model precision, model structure should be reconstructed appropriately. It is the important problem for 3D sensing using cameras in the computer vision

system. It should be noted that the ray tracing approach is not applicable at all in constructing the camera model, unlike CG problem. Although the problem is difficult, we tackle this problem by applying tile coding (CMAC) [7][8] (NOTE: reinforcement learning is not applied in this study), an approach that approximates optical structure of the camera. Our basic concept of the modeling is to reflect locality of relationship between images and real coordinates in depth direction. In this study, considering the modern camera structure with lens practically, we assume that the distortion of lens is negligibly little. In other words, we focus on modeling the optical projection characteristics of a camera in this study.

3.1 Model Structure of Tile Coding

In this study, we consider a state variable expecting to compensate the modeling error of approximation by the conventional model. The state s is the distance between the target and the camera. We propose the camera model based on tile coding structure as:

$$T_{ij} : \omega_{ij} \cdot U = P_{ij} \cdot Q, \quad i = 1,...,m,\ j = 1,...,n \qquad (6)$$

where T_{ij} is the j-th tile (local model) of the i-th layer, ω_{ij} is the scaling parameter of the j-th tile of the i-th layer, P_{ij} is the projection matrix of the j-th tile of the i-th layer, n is the number of tiles in a layer, and m is the number of layers. The conceptual figure of the tile coding is shown in Fig.4. Each tile has a local model with the scaling parameter and projection matrix as in Eq. (6).

Then, we define activation function μ of the tile T_{ij} as:

$$\mu_{ij}(s) = \begin{cases} 1 & ;if\ b_{i,j-1} < s \le b_{ij} \\ 0 & ;else \end{cases} \quad , j=1,\ldots,n \qquad (7)$$

where b_{ij} is the upper limit of the tile T_{ij}, and b_{i0} is the lower limit of the tile T_{i1}. We formulate the tiles by defining the limit b in advance. In general, b is defined to have even width. The tile is utilized to perform processing when the state s is within the range defined in Eq. (7).

In the modeling (calibration) phase, each parameters of each tile should be estimated using calibration data which are pairs of 3D actual point and corresponding image 2D point. We perform modeling by estimating parameters (P and ω) for each tile simply based on selected calibration data set whose activation function value is equal to 1. Unlike CMAC, we apply standard modeling technique such as LMSE based only on selected data set instead of learning technique.

When the state is measured or decided to be s*, the parameters are calculated as:

$$\tilde{\omega} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} \mu_{ij}(s*)\cdot\omega_{ij}\right) \Big/ m \qquad (8)$$

$$\tilde{P} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} \mu_{ij}(s*)\cdot P_{ij}\right) \Big/ m \qquad (9)$$

From these parameters, the camera model is specified as:

$$\tilde{\omega}\cdot U = \tilde{P}\cdot Q \qquad (10)$$

In this way, the tile models are unified as a single model. Good interpolation performance according to the position of the target in the depth direction is expected by this structuring of tile coding.

Figure 5 shows the example of the model decision. In the example, from the measured or decided state s*, the necessary parameters, $\tilde{\omega}$ and $\tilde{P}$, are determined by simple averaging as in Eq.(8) and Eq.(9).

3.2 Stereo Vision Application of Tile Coding Model

We should use generally multiple cameras to measure the 3D coordinates of a target point. In this study, we assume that two cameras (left, right) are installed as the stereo vision and each tile coding model is constructed through calibration in advance of using as the camera model. In order to apply the proposed tile coding model for 3D sensing, the inverse problem of the model should be solved appropriately. In the proposed technique, the inverse problem is to decide the 3D world point from measured 2D image coordinates. In the conventional stereo vision system, the 3D point is calculated easily by using two 2D image coordinates from stereo cameras (left, right). As for the proposed model, the inverse problem is not easily solved according to the model structure. Firstly, this is because the proposed model includes unknown state s, i.e., the distance between the target and the camera, which is the result of the measurement. To solve the problem, we use the conventional model of Eq. (4) to obtain the initial value of the distance, along with the proposed tile coding model. Then using the distance, the target 3D point is calculated based on the tile coding models and iteratively calculated until the target 3D point is converged. Then, from the model parameters and the measured image pixel coordinates of the left and right cameras, we can obtain the 3D point coordinates by using the pseudo-inverse matrix technique as like the conventional manner.

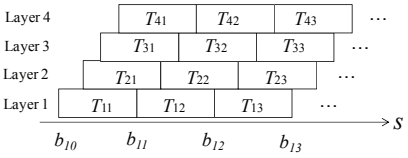

Figure 4. Tile coding structure

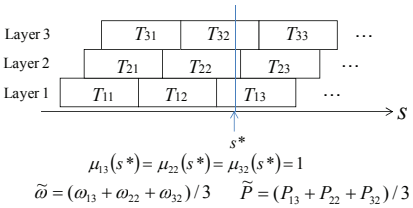

Figure 5. Tile coding example

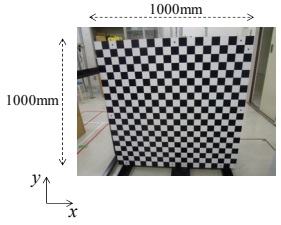

Figure 6. Calibrator

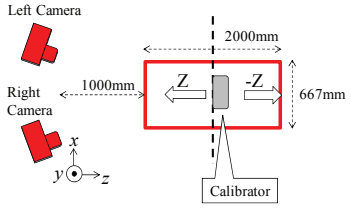

Figure 7. Test field

Table 1. Results ($n = 6$)

	Left$_{[pixel]}$	Right$_{[pixel]}$	Stereo$_{[mm]}$
Conventional Model AVE	2.090	1.724	4.732
Conventional Model STD	2.134	1.418	3.012
Crisp Partitioned Model AVE	1.323	1.187	2.702
Crisp Partitioned Model STD	1.465	1.160	2.109
Tile Coding (m=2) AVE	1.136	1.014	1.792
Tile Coding (m=2) STD	1.360	1.038	1.625
Tile Coding (m=3) AVE	1.148	1.005	1.854
Tile Coding (m=3) STD	1.497	1.065	1.801

Table 2. Results ($n = 8$)

	Left $_{[pixel]}$	Right$_{[pixel]}$	Stereo$_{[mm]}$
Conventional Model AVE	2.090	1.724	4.732
Conventional Model STD	2.134	1.418	3.012
Crisp Partitioned Model AVE	1.145	1.050	2.282
Crisp Partitioned Model STD	1.573	1.120	2.035
Tile Coding (m=2) AVE	1.083	0.969	2.121
Tile Coding (m=2) STD	1.576	1.089	2.077
Tile Coding (m=3) AVE	1.027	0.947	1.935
Tile Coding (m=3) STD	1.408	1.030	1.755

4. EXPERIMENT

We conduct the 3D sensing experiments based on the stereo vision using two industrial cameras. The problem is to decide 3D coordinates of target points precisely from measured 2D image pixel coordinates of left and right cameras. The cameras are high performance megapixel CMOS models with Gigabit Ethernet connection to host PC. The resolution of the camera is 2592x1944 pixels. The lenses are aspheric with less distortion. Since the lens distortion is comparatively little, we do not apply any compensation techniques for the lens distortion in this experiment. The checkerboard calibrator as shown in Fig.6 is used for collecting data pairs for calibration. We evaluate the accuracy of the proposed tile coding model. In this study, prediction accuracy of the camera model is evaluated based on the 2-fold cross validation. The evaluation is performed as follows. The data for calibration are collected at the experiment field as shown in Fig. 7. In the field, two cameras are fixed and the calibrator is moved gradually to collect the calibration data uniformly in the field space. Each data consists of a pair of 2D image coordinates of left and right cameras and corresponding exact 3D world coordinates of the target, which is the corner point of the checker on the calibrator. The number of collected data is 3150. The collected data are split into data set A and data set B. The number of each data set is 1575. The model parameters are estimated using data set A and the accuracy of the identified model is evaluated using data set B. Conversely, the model parameters are also estimated using data set B and the accuracy of the identified model is evaluated using data set A. We evaluate 3D sensing performance of the model using averaged absolute error and standard deviation of the prediction.

We evaluate the proposed tile coding model changing the number of tiles in the layer and the number of layers compared with the conventional model by Open CV [4, 5] and crisp partitioned model [9] which is equivalent to the 1-layered tile coding model. The results are shown in Table 1 and Table 2. Table 1 shows the results of 6 tiles in the layer. Table 2 shows the results of 8 tiles in the layer. The width of tiles is set as even width in the layer. The performance is evaluated as the average value of the absolute prediction error (AVE) and the standard division of prediction error (STD).

4.1 Modeling Results of Camera Model

The sole camera model performance is evaluated based on the precision of the image pixel predicted from 3D target point by the estimated model. The proposed model outperforms the conventional model and crisp partitioned model. As for the crisp partitioned model, the performance tends to be deteriorated according to the number of partitions. This is because the sparsity of calibration data affects modeling performance of the model.

4.2 Results of Stereo Vision

We see from the results in Table 1 and 2 show that the proposed tile coding model outperforms the conventional models. The performance by the proposed model tends to be improved by increasing the number of tiles and the number of layers. However excessive number of tiles might lead to instable performance like crisp partitioned method. The appropriate number of tiles and layers should be decided. We consider that the reason why the proposed model attains good performance is to construct the camera model appropriately reflecting the optical projection characteristics by using tile coding structure on the depth direction.

Moreover, the proposed tile coding model has simple structure and good performance. It can be said that the proposed tile coding model is promising to model the camera and/or projector in the computer vision system.

5. CONCLUSION

In this paper, we proposed a modeling approach based on tile coding for 3D sensing utilizing the configuration of the stereo vision. A distance between a sensing target and a camera is used for structuring the tile coding model for cameras in order to improve approximation performance of camera model in the viewpoint of essential projection characteristics. Through sensing experiments of stereo vision based on the proposed approach, we showed a better performance of our tile based approach compared to the crisp [9] or conventional method [4].

Our future works include evaluation for various cameras and lenses such as inexpensive cameras and application to various measurement problems.

6. REFERENCES

[1] Hartly, R. and Zisserman, A. 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

[2] Deguchi, K. 2000. *Foundation of Robot Vision*. Corona Publishing.

[3] Sato, J. 1999. *Computer Vision – Geometry of Vision –*. Corona Publishing.

[4] Zhang, Z. 2000. A Flexible New Technique for Camera Calibration. *IEEE Trans. on PAMI*. 1330-1334.

[5] Zhang, Z. 1998. *A Flexible New Technique for Camera Calibration. Microsoft Technical Report*. MSR-TR-98-71.

[6] Fujigaki, M. 2009. Whole-Space Tabulation Method for Real-time Shape Measurement and Compact Strain Distribution Measurement System. In *Proc. of ICCES'09*. 566-566.

[7] Miller, W. T. and Parks, P. C. 1991. Design Improvements in Associative Memories for Cerebellar Model Articulation Controllers. In *Proc. ICANN*. 1207-1210.

[8] Sutton, R. S. and Barto, A. G. 1998. *Reinforcement Learning: An Introduction*. MIT Press.

[9] Saito, Y. and Watanabe, T. 2012. A 3D Sensing Technique Using Fuzzy Modeling Based on Stereo Vision. In *Proc. of The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligent Systems*. 2138-2142.